Group of Governmental Experts of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects

10 November 2017

English only

Geneva, 13–17 November 2017
Item 6 of the revised provisional agenda
Examination of various dimensions of emerging technologies in the area of lethal autonomous weapons systems, in the context of the objectives and purposes of the Convention

A "compliance-based" approach to Autonomous Weapon Systems

Working Paper submitted by Switzerland[1]

I. Introduction

1. Ever greater investments go into research and development in the area of robotics and artificial intelligence. This trend is likely to bring increasingly autonomous capabilities to weapon systems, which in turn bears the potential of redefining the role of humans in the use of force.

2. Switzerland is convinced that compliance with international law, notably international humanitarian law (IHL), must be central to discussions about autonomous weapon systems and should figure prominently in the report of the GGE as well as in the way forward.

3. The present working paper recapitulates requirements for compliance with IHL, reiterates considerations by Switzerland on a possible working definition of autonomous weapon systems, and on that basis, identifies elements for a "compliance-based" approach aimed at advancing the debate within the CCW in an inclusive and constructive manner.

II. Ensuring compliance with international humanitarian law

4. IHL is the most relevant body of international law governing the development of autonomous weapon systems and their employment in armed conflicts and will therefore be the main focus of this section. This being said, other branches of international law, most notably human rights law, may equally impose limits on the use of force in armed conflicts, and international criminal law governs individual criminal responsibility for violations. Moreover, any recourse to the threat or use of force by States is governed by the UN Charter.

[1] Adapted version of the working paper submitted to the Informal meeting of experts on lethal autonomous weapons systems (LAWS), Geneva, 11-15 April 2016. Changes were made with regard to the working definition in paragraphs 26 – 29 and the process in paragraph 37.

5. There is general consensus that the potential development and employment of any autonomous weapon systems must remain in compliance with existing international law and, in times of armed conflict, particularly with IHL. The International Court of Justice was clear in its 1996 Advisory Opinion that the established principles and rules of humanitarian law applicable in armed conflict apply to "all forms of warfare and to all kinds of weapons, those of the past, those of the present and those of the future".[2] Even though the treaty provisions regulating the conduct of hostilities do not expressly refer to new technologies of warfare as such, the development and employment of such technologies in armed conflict always must be in accordance with IHL. Customary IHL rules, in particular those related to the conduct of hostilities, equally apply to all means and methods of warfare. Indeed, it is a long standing principle that the right of parties to an armed conflict to choose methods or means of warfare is not unlimited.

6. Against this background, three distinct issues should be considered: (1) the substantive provisions of IHL applicable to the employment of autonomous weapon systems in armed conflict; (2) the corresponding procedural rule of IHL requiring the conduct of legal weapons reviews as a means of ensuring conformity with international law; and (3) the issue of accountability.

A. Substantive IHL provisions

7. At the outset, it is important to distinguish between the lawfulness of a particular type of weapon as such (weapons law), and the lawfulness of the way in which it is being used (targeting law). While every weapon can be used in an unlawful manner, the inherent characteristics of certain weapons categories entail that their use – in some or all circumstances – is unlawful *per se*. The permissibility of all other weapons depends on their employment in conformity with IHL.

8. Under IHL, any weapon possessing one or more of the following characteristics is inherently unlawful : (1) the weapon is of a nature to cause superfluous injury or unnecessary suffering; (2) the weapon is indiscriminate by nature because it cannot be aimed at a lawful target or because its effects cannot be limited as required by IHL; (3) the weapon is intended, or may be expected, to cause widespread, long-term and severe damage to the natural environment; (4) the weapon has been specifically prohibited in treaty or customary law. These criteria apply to all weapons, including new technologies such as autonomous weapon systems.

9. Of particular relevance for autonomous weapon systems is the prohibition of indiscriminate weapons. A weapon system would have to be regarded as indiscriminate if it cannot be directed at a specific military objective or if its effects cannot be limited as required by IHL and if, in either case, it is of a nature to strike military objectives and civilians or civilian objects without distinction. In other words, in order for an autonomous weapon system to be lawful under this rule, it must be possible to ensure that its operation will not result in unlawful outcomes with respect to the principle of distinction.

10. With regard to the *lawful use* of a weapon system, the principles governing the conduct of hostilities need to be considered. Most notably, in order to lawfully use an autonomous weapon system for the purpose of attack, belligerents must: (1 - Distinction) distinguish between military objectives and civilians or civilian objects and, in case of doubt, presume civilian status; (2 - Proportionality) evaluate whether the incidental harm likely to be inflicted on the civilian population or civilian objects would be excessive in relation to the concrete and direct military advantage anticipated from that particular attack; (3 - Precaution) take all feasible precautions to avoid, and in any event minimize, incidental harm to civilians and damage to civilian objects; and cancel or suspend the attack if it becomes apparent that the target is not a military objective, or that the attack may be expected to result in excessive incidental harm.

[2] International Court of Justice, Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion of 8 July 1996, para. 86.

11. The employment of autonomous weapon systems in the conduct of hostilities also raises particular challenges with regard to the prohibition of the denial of quarter and the protection of persons *hors de combat*, i.e. the protection from attack of the wounded and sick and those intending to surrender.[3] Any reliance on autonomous weapon systems would need to preserve a reasonable possibility for adversaries to surrender. A general denial of this possibility would violate the prohibition of ordering that there shall be no survivors or of conducting hostilities on this basis (denial of quarter).[4]

12. Full compliance with IHL, however, is not limited to the rules governing the conduct of hostilities. Besides employing autonomous weapon systems as a weapon in attack, it is also conceivable that such systems could be used to perform other tasks governed by IHL, such as the guarding and transport of persons deprived of their liberty or tasks related to crowd control and public security in occupied territories. Additional specific rules need to be taken into consideration if autonomous weapon systems were to be relied on for such activities

13. Applying these requirements of *lawful use* to autonomous weapon systems is not without complexity. The Geneva Conventions of 1949 and their Additional Protocols of 1977 were undoubtedly conceived with States and individual humans as agents for the exercise and implementation of the resulting rights and obligations in mind. In addition, many pivotal rules of IHL presume the application of evaluative decisions and value judgements, such as the presumption of civilian status in case of "doubt",[5] the assessment of "excessiveness" of expected incidental harm in relation to anticipated military advantage, the betrayal of "confidence" in IHL in relation to the prohibition of perfidy, and the prohibition of destruction of civilian property except where "imperatively" demanded by the necessities of war.[6] The principle of precaution even expressly refers to "those who plan or decide upon" an attack,[7] and the provisions establishing criminal responsibility for serious violations of IHL also are based on a manifest presumption of human agency.[8]

14. Furthermore, the overarching obligation of all belligerents to "respect and ensure respect" for IHL "in all circumstances" seems to imply a derived duty of exercising sufficient control or supervision over the development and/or employment of autonomous weapon systems to ensure full compliance with IHL and prevent outcomes that would be unlawful under existing international law. In accordance with this obligation, it is uncontested that preparatory measures must be taken to permit the implementation of IHL and implementation should be supervised.[9]

15. Accordingly, given the current state of robotics and artificial intelligence, it is difficult today to conceive of an autonomous weapon system that would be capable of reliably operating in full compliance with *all* the obligations arising from existing IHL without any human control in the use of force, notably in the targeting cycle.

16. On this basis, the question, therefore, is not whether States have a duty to control or supervise the development and/or employment of autonomous weapon systems, but how that control or supervision ought to be usefully defined and exerted. Would it be sufficient, for example, to rely on superior programming and strict reliability testing to make an autonomous weapon system predictably compliant with IHL for its intended operational

---

[3] Article 41 of Additional Protocol I to the Geneva Conventions.

[4] Article 40 of Additional Protocol I to the Geneva Conventions.

[5] Article 50(3) and article 52(3) of Additional Protocol I to the Geneva Conventions.

[6] Article 23(g) of Hague Regulation IV, Article 53 of the Fourth Geneva Convention.

[7] Article 57(2)(a) of Additional Protocol I to the Geneva Conventions.

[8] Article 49/50/129/146 Geneva Conventions; Section III of Additional Protocol I to the Geneva Conventions.

[9] Cf. Yves Sandoz, Christophe Swinarski, Bruno Zimmermann (eds.), Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949, ICRC, 1987, Article 1 Additional Protocol I, para 41. See further article 80 of Additional Protocol I to the Geneva Conventions, which requires that the High Contracting Parties and the parties to the conflict shall "take all necessary measures for the execution of their obligations", "give orders and instructions to ensure observance" and "supervise their execution".

parameters? If so, would it be permissible to restrict human involvement to the proper activation of such an autonomous weapon system? This working paper does not seek to prejudge these questions.

17. However, it is useful to recognize that control can be exercised in various different ways, both independently and in combination. Arguably, in the future, a significant level of control can already be exerted in the development and programming phase. Through testing and evaluating autonomous weapon systems in the course of weapons reviews, predictability and reliability of such systems can also be reinforced. Predictability and reliability can also be increased by restricting the autonomous weapon systems' parameters of engagement in line with the system's compliance capabilities. Depending on operational requirements and system capabilities, further control can be exercised through real-time supervision, or through an autonomous or human operated override mechanism aimed at avoiding malfunction or, alternatively, ensuring safe failure. Assessing, amongst others, these parameters in relation with existing weapon systems may contribute to a better understanding of appropriate means of ensuring compliance with IHL in the development and use of autonomous weapon systems for military purposes. As an illustration, those current weapon systems that are expected to operate with limited human control, such as certain booby-traps, are generally subject to severe restrictions as to the geographic area and mode of their use. The applicability of such restrictions with regard to systems with far greater and more sophisticated autonomy could be assessed in order to understand their relevance in terms of IHL compliance.

18. This overview of the relevant rules of IHL would be incomplete without reference to the Martens clause. The Martens Clause, which is part of customary international law, affords an important fallback protection in as much as the "laws of humanity and the requirements of the public conscience" need to be referred to if IHL is not sufficiently precise or rigorous. [10] Accordingly, not everything that is not explicitly prohibited can be said to be legal if it would run counter the principles put forward in the Martens clause. Indeed, the Martens clause may be said to imply positive obligations where contemplated military action would result in untenable humanitarian consequences.

B. Legal reviews

19. Under IHL, the substantive rules described above are complemented by a procedural rule. As with any other weapon, means or method of warfare, States have the positive obligation to determine, in the study, development, acquisition or adoption of any autonomous weapon system, whether their employment would, in some or all circumstances, contravene existing international law. In this regard the duty to conduct legal reviews, as specified in article 36 of Additional Protocol I to the Geneva Conventions, constitutes an important element in preventing or restricting the development and employment of new weapons that would not meet the obligations listed above.[11] Moreover, adequate testing and reviews may also have implications on the level of State responsibility, including for malfunction of approved autonomous weapon systems.

20. The legal review of autonomous weapon systems may present a number of challenges distinct from traditional weapons reviews. Specifically, the question is how such

[10] The "principles of humanity and the dictates of public conscience" are contained in the so called Martens Clause, which is repeated notably in the preamble of the CCW and in Art. 1(2) of Additional Protocol I. In its 1996 Advisory Opinion on the legality of the threat or use of nuclear weapons, the International Court of Justice held that the clause "proved to be an effective means of addressing the rapid evolution of military technology" (§78).

[11] ICRC, A Guide to the Legal Review of New Weapons, Means and Methods of Warfare: Measures to Implement Article 36 of Additional Protocol I of 1977 (2006), available at https://www.icrc.org/eng/assets/files/other/icrc_002_0902.pdf. See also the working paper submitted by the Netherlands and Switzerland to the 2017 Group of Governmental Experts on Lethal Autonomous Weapons Systems (LAWS), Weapons Review Mechanisms (CCW/GGE.1/2017/WP.5), available at: http://undocs.org/ccw/gge.1/2017/WP.5.

systems and their specific characteristics can be meaningfully tested. Beyond the purely technical challenge of assessing IHL compliance of an autonomous weapon system, there is also a conceptual challenge related to the fact that an autonomous system will assume an increasing number of determinations in the targeting cycle which traditionally are being taken care of by a human operator. For example, in traditional systems, the principle of proportionality was to be respected by the operator. It consequently fell outside the scope of an article 36 review. However, if an autonomous weapon system is expected to perform this proportionality assessment by itself, that aspect will need to be added to legal reviews of these systems. New evaluation and testing procedures may need to be conceptualized and developed to meet this particular challenge.

21. Given the special characteristics of autonomous weapon systems, a number of further measures could be recommended for incorporation into national review procedures. For example, one could imagine recommending that in some cases, particular safeguards against malfunction, such as the possibility of a human override, are built into autonomous weapon systems. Proper understanding of a system's predictability, especially when it comes to interaction with other autonomous systems, could also be named as example. While the process of national legal reviews may require procedural and technical adaptations to fully capture the complexity of autonomous weapon systems, if rigorously implemented, it holds the potential of ensuring that all new weapons, means and methods of warfare are developed and acquired in compliance with international law.

C. Accountability

22. Another important issue arising with regard to autonomous weapon systems is that of accountability, namely in terms of individual criminal responsibility and of state responsibility. Given that autonomous weapon systems possess no agency or legal personality of their own, the question of individual criminal responsibility focuses entirely on the responsibility of humans that are involved as operators, commanding officers, programmers, engineers, technicians or in other relevant functions. If the deployment of an autonomous weapon system results in a serious violation of IHL, and if that violation is the consequences of culpable fault on the part of a human being the latter may be subjected to criminal prosecution for war crimes or, depending on the circumstances of the case, also for crimes against humanity or genocide.

23. Criminal culpability is self-evident in the case of deliberate and premeditated intent. It is less so in the case of recklessness or (advertent) negligence, or of simple acceptance of a risk that violations will or may occur. With regard to war crimes, article 85(3) of Additional Protocol I requires "willfulness", with national practices varying as to the meaning to be given to this requirement. The International Tribunal for the Former Yugoslavia has stated that, as a matter of customary law, indirect intent would be sufficient to fulfil the mental requirement *(mens rea)*.[12] Conversely, the Rome Statute of the International Criminal Court does not foresee criminal liability for indirect intent,[13] except in the case of command responsibility for the conduct of subordinates.[14] As a matter of concept, command responsibility does not entail the commander's direct criminal responsibility for crimes committed by his subordinates, but for his or her culpable failure to prevent, suppress or repress crimes committed by persons (i.e. not machines) under his or her command and control. Strictly speaking, therefore, a commander's failure to duly control autonomous weapon systems operating under his command is not a case of command responsibility within the contemporary understanding of this concept, but may constitute a direct violation of the duties of precaution, distinction, proportionality or any other obligation imposed by IHL. This does not exclude that, as the functions of human soldiers are increasingly "delegated" to autonomous weapon systems, it may become appropriate *de lege ferenda* to extend the commander's supervisory duty, *mutatis mutandis*

[12] Cf. ICTY, Prosecutor v. Tihomir Blaskic, Judgement of 29 July 2004, Appeals Chamber, para. 42.
[13] Article 30 of the Rome Statute of the International Criminal Court.
[14] Article 28 of the Rome Statute of the International Criminal Court.

and *by analogy*, also to autonomous weapon systems operating under his direct command and control.

24. Overall, under current international law, whether or not there is an "accountability gap" for operators, commanders and other humans involved in the operation of autonomous weapon systems depends on the applicable *mens rea* standard. As a general assumption, the more significant human involvement in a specific autonomous weapon system operation is (such as humans "in the loop"), the easier it is to assign individual responsibility. This assumption may be relevant with a view to the general obligation of States to respect and ensure respect for IHL.

25. The second dimension of accountability derives from general international law governing the responsibility of States for internationally wrongful acts. States remain legally responsible for unlawful acts and resulting harm caused by autonomous weapon systems they employ, including due to malfunction or other undesired or unexpected outcomes. The rules governing *attribution* of conduct to a State are pertinent in relation to autonomous weapon systems as with any other means and methods of warfare. Given that autonomous weapon systems lack legal personality in the first place, they cannot become agents in a human sense, whether state agents or non-state actors. The question of State responsibility therefore does not turn on the nature or capability of the autonomous weapon systems, but of legal and factual status of the person or entity deciding on its employment. A decision of a person or entity exercising public powers or governmental authority (e.g. the armed forces) to employ an autonomous weapon system in a given situation certainly would be attributable to the State.[15] The result is that States cannot escape international responsibility by a process of "delegating" certain tasks to autonomous weapon systems.

III. Working definition

26. Switzerland remains of the view that, at the current stage, it is premature to aim for a *definition* of (lethal) autonomous weapon systems that seeks to draw a line between desirable, acceptable or unacceptable systems. However, Switzerland considers it relevant that the GGE endeavours to develop a common understanding of both the substance and the purpose of a *preliminary working definition*.

27. With regard to substance, such a working definition should provide for a shared understanding that there is a need to focus on systems with elements of autonomy in the targeting cycle. In addition, this working paper considers that the element of lethality, though of particular concern in practice, should not be conceptually regarded as a prerequisite characteristic of autonomous weapon systems. It prefers a more inclusive understanding of autonomous weapon systems, which would also cover means and methods of warfare that do not necessarily inflict physical death, but the effects of which may be restricted to causing, for example: (1) physical injury short of death, (2) physical destruction of objects, or (3) non-kinetic effects.

28. With regard to the purpose of the working definition, Switzerland is of the view that such a working definition should not prejudge the question of which systems may potentially require a regulatory response. Keeping in mind that fully autonomous weapon systems do not yet exist, it should allow for the examination of all relevant systems or capabilities including existing ones. With a view to ensuring compliance with international law – the very topic of this paper – such an examination may contribute to our understanding of the challenges that autonomy in a weapon system may raise and permit the identification of practical measures, best practices and standards.

29. Against this background, Switzerland suggested in 2016 to describe autonomous weapon systems as *"weapons systems that are capable of carrying out tasks governed by IHL in partial or full replacement of a human in the use of force, notably in the targeting cycle"*. Such a working definition is inclusive, accounts for a wide array of system

[15] Article 4 et seq. of the Articles on State Responsibility for Internationally Wrongful Acts (2001). See also Article 91 of Additional Protocol I to the Geneva Conventions.

configurations, and allows for a debate that is differentiated, compliance-based, and without prejudice to the question of appropriate regulatory response. Indeed, the proposed working definition is not conceived in any way to single out only those systems which could be seen as legally objectionable or otherwise requiring regulation. At one end of the spectrum of systems falling within that working definition, States may find some subcategories to be entirely unproblematic, while at the other end of the spectrum, States may find other subcategories unacceptable. As discussions advance, this working definition needs to evolve to become more specific and purposeful.

IV. A compliance-based approach

30. The present working paper has sought to map the most relevant IHL obligations applicable to the development and employment of autonomous weapon systems on the basis of a broad, inclusive working definition that allows for a discussion of different types of increasingly autonomous weapon systems. On one end of the spectrum of the proposed working definition, some types of autonomous weapon systems would be already unlawful under existing IHL, while on the other end of the spectrum, some types of autonomous weapon systems can be readily qualified as unproblematic.

31. Given the consensus that existing international law applies to all weapon systems, including autonomous weapon systems, and that it has to be respected in all circumstances, an IHL "compliance-based" approach presents itself as a way forward for the CCW discussions. The approach would have three main parts:

32. Firstly, in order to advance our understanding of autonomy and its relation to IHL, States could, based on the proposed working definition, *assess existing autonomous weapon systems and existing systems with* limited autonomy in the targeting cycle. The specific parameters which make a particular system IHL compliant could be identified and examined. These could then be extrapolated to future systems with higher levels of autonomy, to gain an understanding of what features contribute to conformity with or – conversely – objectionability under IHL.

33. Secondly, *reaffirm and spell out applicable international law, in particular IHL*. While many provisions of IHL are well known in principle, there would seem to be merit in collating and clarifying, for ease of reference, the relevant existing provisions as they apply to autonomous weapon systems . Such an exercise would involve the three angles treated in this working paper, i.e. the substantive IHL provisions, legal reviews, and accountability. This working paper has also put forward the notion that – given the current state of robotics and artificial intelligence – the relevant question is not whether a certain level of human control is called for, but what kind and level of human involvement in each of the different phases ranging from conceptualization, development and testing, to operational programming, employment and target engagement. At the heart of the issue is the question: what is the right quality of the human-machine interaction to ensure and facilitate compliance with IHL?

34. Thirdly, still with a view to securing and facilitating compliance, *identify best practices, technical standards and policy measures* that, rather than being strictly part of applicable international obligations, complement, promote and reinforce them. For example, with regard to legal reviews, it could be of interest to identify standard methods and protocols for testing autonomous weapon systems that take into account their unique characteristics. In the same vein, should gaps in the chain of accountability become apparent, States would potentially want to discuss complementary or additional means of ensuring that individual accountability is maintained.

35. The above is not meant to exhaust the debate and is without prejudice to *contemplating further regulatory responses* as may be deemed required. Indeed, discomfort has been expressed by several States and civil society organizations that the delegation of "life and death" decisions to machines is unacceptable and would run against the principles of humanity and the dictates of public conscience. This highlights the importance of human dignity as one of the core values of our international community and as a key element underpinning many international instruments, notably the sources of IHL. While the

question whether there should be limits to handing over certain functions to autonomous weapon systems is a legitimate and important discussion that has its place, it is a question that is very different from the one of compliance with existing international law.

36. In this sense, the "compliance-based" approach submitted here is meant to be part and parcel of a broader "building block" approach that gives space for legal, military, ethical and other legitimate considerations to inform discussions on possible regulatory responses. There appears to be a continued interest in addressing all relevant aspects relating to autonomous weapon systems, notably technological definitions, military utility as well as legal and ethical aspects, under involvement of a variety of stakeholders.

37. With regard to CCW process itself, the work of the Governmental Group of Experts (GGE) provides the opportun y to consolidate the broadly shared understanding that the application of, and compliance with, International Law and in particular International Humanitarian Law. There would be merit to continue studying, based on a broad working definition, all aspects related to the increasing autonomy in weapon systems and, in accordance with the approach presented in this paper, include work along the three main areas of activity mentioned above.